Learning by a neural net in a noisy environment – The pseudo-inverse solution revisited

W A van Leeuwen and B Wemmenhove†

Institute for Theoretical Physics, University of Amsterdam, Valckenierstraat 65, 1018 XE Amsterdam, The Netherlands

Abstract. A recurrent neural net is described that learns a set of patterns $\{\boldsymbol{\xi}^\mu\}$ in the presence of noise. The learning rule is of a Hebbian type, and, if noise would be absent during the learning process, the resulting final values of the weights w_{ij} would correspond to the pseudo-inverse solution of the fixed point equation in question. For a non-vanishing noise parameter, an explicit expression for the expectation value of the weights is obtained. This result turns out to be unequal to the pseudo-inverse solution. Furthermore, the stability properties of the system are discussed.

Submitted to: *J. Phys. A: Math. Gen.*

PACS numbers: 84.35+i, 87.10+e, 87.18.Sn, 07.05.Mh

† Correspondence should be addressed to wemmenho@science.uva.nl

1. Introduction and summary

In principle, it is our purpose to study learning in a neural net as it occurs in nature. The theory of recurrent neural nets [1] provides us with a model of content-addressable memory as it might be realized, to some extent, in the brain. Learning, in such a model, corresponds to adjusting the synaptic matrix w_{ij} in such a way that p memorized patterns $\boldsymbol{\xi}^\mu$, $(\mu = 1, \ldots, p)$, become fixed points of the neuron state dynamics. This can be achieved in a recurrent neural net by sequentially clamping its neurons to a well-defined and unique set of patterns, and adjusting the weights of the connections according to some Hebbian learning rule. However, in reality, a neural net cannot be clamped to a fixed set of ideal patterns. A more realistic assumption would be that the clamping of the net to a pattern always is more or less distorted. Consider, for instance, the visual system as a system in which the clamping is imposed by input from the retina. Since neurons are noisy objects, which once in a while fire spontaneously, an internal representation of a stimulus in the brain will hardly ever be identical to the representation corresponding to a previous stimulus.

We therefore introduce noise to the set of patterns, thus making the set less well-defined and less unique. A network state array of a net of N neurons is denoted by

$$\boldsymbol{x} := (x_1, \ldots, x_N) \tag{1}$$

where $x_i = 1$ if neuron i is active and $x_i = 0$ if it is non-active. At every learning step n, $\boldsymbol{x}$ will be similar to one of the p given patterns $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p$, but it has nonzero probability, for each neuron i, $(i = 1, \ldots, N)$, of deviating from it. At each learning step n, synaptic connections w_{ij} will adapt themselves, according to a Hebbian learning rule which is a function of the weights w_{ik} and of the (binary) neuron states x_k, $(k = 1, \ldots, N)$. For the class of learning rules that we will use in our model, the case of noiseless learning has been studied in detail ([2], [3], [4]). At every learning step n a pattern $\boldsymbol{\xi}^\mu$, $(\mu = 1, \ldots p)$ is chosen. If, for each n, we put $\boldsymbol{x} = \boldsymbol{\xi}^\mu$, the resulting weights w_{ij} for $n \to \infty$ are known to coincide with the pseudo-inverse solution. The pseudo-inverse solution is a particular solution of the under-determined set of pN equations

$$\sum_{j=1}^{N} w_{ij}\xi_j^\mu - \theta_i = \kappa(2\xi_i^\mu - 1) \tag{2}$$

for the $N(N-1)$ unknowns w_{ij} $(i, j = 1, \ldots, N; i \neq j)$, $p < N$. Here, κ is a positive number, and θ_i a constant. It is easily verified that these equations guarantee that the so-called stability coefficients

$$\gamma_i(\boldsymbol{\xi}^\mu) = (2\xi_i^\mu - 1)\left(\sum_{j=1}^{N} w_{ij}\xi_j^\mu - \theta_i\right) \tag{3}$$

are positive for all patterns $\boldsymbol{\xi}^\mu$. In our description, however, the network state during learning is determined by a probability distribution $p^\mu(\boldsymbol{x})$, centered around the patterns $\boldsymbol{\xi}^\mu$. By means of the Master equation derived in section 2, we will arrive at the following equation for the expectation value of the weights in the limit of $n \to \infty$:

$$\frac{1}{p}\sum_{\mu=1}^{p}\sum_{\boldsymbol{x}\in\Omega} p^\mu(\boldsymbol{x})[\kappa(2x_i - 1) - (\sum_{k=1}^{N}\langle w_{ik}\rangle_\infty x_k - \theta_i)]x_j = 0 \tag{4}$$

where Ω is the collection of all possible 2^N arrays $\boldsymbol{x}$. In contrast to the equations (2), this is a completely determined set of equations, of which the solution is essentially different from the pseudo-inverse solution of (2). It will turn out to exist only if the variance of the probability distribution $p^\mu(\boldsymbol{x})$ is non-zero, *i.e.*, in the presence of noise. This is the subject of section 3.

In section 4 we will study numerically the weights $\langle w_{ij} \rangle_n$ as a function of n.

In section 5 we will show that, on the average, this solution does yield stability coefficients close to κ if the 'noise parameter', b, which will be introduced in the probability distribution $p^\mu(\boldsymbol{x})$, is small enough, which shows that the solution found is stable indeed.

Finally, in section 6, we study the size of the basins of attraction around our new solution. It is found that the sizes are larger than those around the pseudo-inverse solution, a result that is in perfect agreement with earlier observations [5, 6, 7, 8] that learning with noise enlarges the basins of attraction. In these earlier studies, however, no analytical expression for the average values of the weights has been given.

2. Derivation of the Master Equation for a linear learning rule

We consider a recurrent net of N binary neurons. The strengths of the synaptic connection between the post-synaptic neuron j and the pre-synaptic neuron i will be denoted by w_{ij}. The neurons i $(i = 1, \ldots, N)$ can take the values $x_i = 0$ or $x_i = 1$, corresponding to the non-active and active state, respectively. It is useful to associate with each neuron i a set V_i, defined as the collection of neuron indices j with which neuron i has an adaptable, *i.e.*, a non-zero, non-constant afferent synaptic connection. In other words, for all $j \in V_i$, there is an axon going from neuron j to a dendrite of neuron i, and the corresponding weight w_{ij} is adaptable in a learning process. The collection of neurons j with which i has no connection, or a non-changing synaptic connection, will be denoted as the complementary set, V_i^C. We suppose that the synaptic strengths w_{ij} between the neurons are changed in steps according to a rule of the general form

$$w'_{ij} = w_{ij} + \Delta w_{ij} \qquad j \in V_i \tag{5}$$

$$w'_{ij} = w_{ij} \qquad j \in V_i^C \tag{6}$$

where Δw_{ij} is a function of the states x_k of all N neurons of the net and all afferent synaptic weights w_{ik} (i fixed, $k = 1, 2, \ldots, N$). In general, the functions Δw_{ij} will be linear in all x_k, since $x_k^2 = x_k$ (recall that x_k equals 0 or 1), but non-linear in the weights w_{ik}. In this article we suppose, however, that the Δw_{ij} do depend linearly on the weights w_{ik}. Hence, in this article,

$$\Delta w_{ij} = \Delta w_{ij}(\boldsymbol{x}, \boldsymbol{w}_i) \tag{7}$$

is a linear function in all x_k and all w_{ik} $(k = 1, 2, \ldots N)$. We abbreviated

$$\boldsymbol{w}_i := (w_{i1}, \ldots, w_{iN}) \tag{8}$$

It is unrealistic to describe a biological neural net as a deterministic system, since there are many unknown parameters that influence its development in time. We therefore choose a probabilistic description. We suppose that the neuron states x_i are mutually independent stochastic variables, *i.e.*, the probability that neuron i has value x_i is given by a probability distribution $p_i(x_i)$ which is independent of j $(j \neq i)$. Since the changes Δw_{ij} of the weights w_{ij} are functions of the stochastic variables x_i

$(i = 1, 2, \ldots N)$, and a function of a stochastic variable is a stochastic variable, the changes Δw_{ij}, and, hence, the w_{ij} themselves, are stochastic variables.

Now let $T_{ij}(w'_{ij}|w_{ij}, \{w_{ik}\}_{k\neq j})$ be the probability that, due to a learning step, a transition takes place from the value w_{ij} to the value $w'_{ij} = w_{ij} + \Delta w_{ij}$, for a given set $\{w_{ik}\}_{k\neq j}$. Then we have

$$T_{ij}(w'_{ij}|w_{ij}, \{w_{ik}\}_{k\neq j}) = \sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x})\delta(w'_{ij} - w_{ij} - \Delta w_{ij}(\boldsymbol{x}, \boldsymbol{w}_i)) \tag{9}$$

where Ω is the collection of all 2^N possible states of the neural net $(x_i = 0, 1; i = 1, \ldots, N)$ and $p(\boldsymbol{x})$ is the probability of occurrence of the network state $\boldsymbol{x}$, which we suppose to be independent of the variables w_{ij}. [The relation between $p(\boldsymbol{x})$ and $p_i(x_i)$ is left unspecified at this stage of the reasoning; compare, however, (26) and (27) below.] We have

$$\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1 \tag{10}$$

The delta-function in (9) guarantees that only transitions take place which obey the learning rule (6). Using (10) we find from (9), that

$$\int T_{ij}(w'_{ij}|w_{ij}, \{w_{ik}\}_{k\neq j})dw'_{ij} = 1 \tag{11}$$

Let $P_{ij}(w_{ij}, n)$ be the probability of occurrence of the variable w_{ij} at a time step n $(n = 0, 1, 2, \ldots)$. Then P_{ij} and T_{ij} are related according to

$$P_{ij}(w_{ij}, n+1) = \int \ldots \int T_{ij}(w_{ij}|\{w'_{ik}\}) \prod_{k=1}^{N} [P_{ik}(w'_{ik}, n)dw'_{ik}] \tag{12}$$

Demanding that the probability P_{ij} is normalized initially,

$$\int P_{ij}(w_{ij}, 0)dw_{ij} = 1 \tag{13}$$

we find from (12) and (11), by induction, that

$$\int P_{ij}(w_{ij}, n)dw_{ij} = 1 \tag{14}$$

for all n. From (11) and (12) it follows that

$$\begin{aligned} P_{ij}(w_{ij}, n+1) - P_{ij}(w_{ij}, n) = &\int \ldots \int [T_{ij}(w_{ij}|\{w'_{ik}\})P_{ij}(w'_{ij}, n) \\ &-T_{ij}(w'_{ij}|w_{ij}, \{w'_{ik}\}_{k\neq j})P_{ij}(w_{ij}, n)] \prod_{k\neq j} [P_{ik}(w'_{ik}, n)dw'_{ik}]dw'_{ij} \\ &(i = 1, \ldots, N; j \in V_i) \end{aligned} \tag{15}$$

which is the so-called Discrete Master Equation for the weights w_{ij}. It masters the evolution of the weights w_{ij} as a function of n, and determines the values of the weights in the long run.

In order to obtain an expression for the expectation value of the weights after infinitely many learning steps, we first consider the expectation value at time step n:

$$\langle w_{ij}\rangle_n := \int P_{ij}(w_{ij}, n)w_{ij}dw_{ij} \qquad (j \in V_i) \tag{16}$$

The latter expression yields, using the Master Equation (15),

$$\langle w_{ij}\rangle_{n+1} - \langle w_{ij}\rangle_n = \int \dots \int w_{ij}[T_{ij}(w_{ij}|\{w'_{ik}\})P_{ij}(w'_{ij},n) \\ -T_{ij}(w'_{ij}|w_{ij},\{w'_{ik}\}_{k\neq j})P_{ij}(w_{ij},n)]\prod_{k\neq j}[P_{ik}(w'_{ik},n)dw'_{ik}]dw'_{ij}dw_{ij} \quad (17)$$

or, interchanging the primed and unprimed variables w_{ij} and w'_{ij} in the first term on the right hand side,

$$\langle w_{ij}\rangle_{n+1} - \langle w_{ij}\rangle_n = \int \dots \int (w'_{ij} - w_{ij})T_{ij}(w'_{ij}|w_{ij},\{w'_{ik}\}_{k\neq j})P_{ij}(w_{ij},n) \\ \times\prod_{k\neq j}[P_{ik}(w'_{ik},n)dw'_{ik}]dw'_{ij}dw_{ij} \quad (18)$$

or, with (9) and integrating over w'_{ij},

$$\langle w_{ij}\rangle_{n+1} - \langle w_{ij}\rangle_n = \sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) \int \dots \int \Delta w_{ij}(\boldsymbol{x},\boldsymbol{w}_i)\prod_{k=1}^{N} P_{ik}(w_{ik},n)dw_{ik} \qquad (j\in V_i)(19)$$

or, with (14) and (16),

$$\langle w_{ij}\rangle_{n+1} - \langle w_{ij}\rangle_n = \sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x})\Delta w_{ij}(\boldsymbol{x},\langle \boldsymbol{w}_i\rangle_n), \qquad (j\in V_i) \qquad (20)$$

where we used that Δw_{ij} is linear in the w_{ik} $(k = 1,\dots,N)$ to replace $\boldsymbol{w}_i$ by the expectation value $\langle \boldsymbol{w}_i\rangle_n$ in the expression for Δw_{ij}. If we assume that the expectation values of the synaptic connections $\langle w_{ij}\rangle_n$ converge to finite values, $\langle w_{ij}\rangle_\infty$, for n tending to infinity, we can solve this equation for $n\to\infty$. This is the subject of the next section.

3. Final values for the weights

If we suppose that the left-hand side of (20) vanishes in the limit of n tending to infinity, we have

$$\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x})\Delta w_{ij}(\boldsymbol{x},\langle \boldsymbol{w}_i\rangle_\infty) = 0 \qquad (21)$$

At this point, we need an expression for the increment $\Delta w_{ij}(n)$, in the n-th learning step. We take the biologically motivated learning rule [4]

$$\Delta w_{ij}(n) = \eta_i[\kappa - \gamma_i(\boldsymbol{x},n)](2x_i - 1)x_j \qquad (i = 1,\dots,N; j\in V_i) \quad (22)$$

where η_i is the learning rate, κ the margin parameter and $\gamma_i(\boldsymbol{x},n)$ the stability coefficient given by

$$\gamma_i(\boldsymbol{x},n) = (2x_i - 1)[h_i(\boldsymbol{x},n) - \theta_i] \qquad (23)$$

[cf. eq. (3)]. Here, $h_i(\boldsymbol{x},n)$ is the membrane potential of neuron i at step n of the learning process

$$h_i(\boldsymbol{x},n) = \sum_{k=1}^{N} w_{ik}(n)x_k \qquad (24)$$

and θ_i the threshold potential of neuron i. It should be noted that in (24) x_k is the state of neuron k at step n of the learning procedure. Substituting (22) with (23) and (24) into (21) we find, using the fact that $(2x_i - 1)^2 = 1$,

$$\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x})[\kappa(2x_i - 1) - (\sum_{k=1}^{N} \langle w_{ik}\rangle_\infty x_k - \theta_i)]x_j = 0 \tag{25}$$

where we divided by the learning rate η_i.

Up to now, the precise form of the probability distribution $p(\boldsymbol{x})$ has been left unspecified. At this point, let us specify our probability distribution $p(\boldsymbol{x})$ to be such that the chosen patterns $\boldsymbol{x}$ are centered around representative patterns $\boldsymbol{\xi}^\mu$. To that end, we choose our probability distribution $p(\boldsymbol{x})$ such that it is a sum of p equally probable, individually independent probability distributions, *i.e.*,

$$p(\boldsymbol{x}) = \frac{1}{p}\sum_{\mu=1}^{p} p^\mu(\boldsymbol{x}) \tag{26}$$

where $p^\mu(\boldsymbol{x})$ is factorizable,

$$p^\mu(\boldsymbol{x}) = \prod_{i=1}^{N} p_i^\mu(x_i) \tag{27}$$

i.e., the neurons behave independently from one another. The quantity $p_i^\mu(x_i)$ is the probability that, once the pattern index μ is chosen, neuron i is in the state x_i. One therefore has

$$p_i^\mu(0) + p_i^\mu(1) = 1 \tag{28}$$

In a learning process, at every step n, the index μ is drawn from a collection of p equally probable pattern indices, thus fixing the probability distribution $p^\mu(\boldsymbol{x})$ according to which the pattern $\boldsymbol{x}$ is chosen for that learning step n.

Let us denote averages with respect to the probability $p^\mu(\boldsymbol{x})$ by $\overline{}^\mu$

$$\sum_{x_i=0,1} p_i^\mu(x_i)x_i = \overline{x_i}^\mu, \qquad \sum_{x_i=0,1} p_i^\mu(x_i)x_i^2 = \overline{(x_i^2)}^\mu \tag{29}$$

implying, in view of (27) and (28),

$$\sum_{\boldsymbol{x}\in\Omega} p^\mu(\boldsymbol{x})x_i = \overline{x_i}^\mu, \qquad \sum_{\boldsymbol{x}\in\Omega} p^\mu(\boldsymbol{x})x_i^2 = \overline{(x_i^2)}^\mu \tag{30}$$

Thus a bar with an index μ indicates an average with respect to the probability distribution $p^\mu(\boldsymbol{x})$. With the choice (26), the result (25) can be rewritten in terms of these averages, where we must take be aware that a term $\overline{(x_j^2)}^\mu$ appears in the sum over k:

$$\langle w_{ij}\rangle_\infty \sum_{\mu=1}^{p} \left[\overline{(x_j^2)}^\mu - (\overline{x_j}^\mu)^2\right] = \sum_{\mu=1}^{p} \left[\kappa(2\overline{x_i}^\mu - 1) - (\sum_{k=1}^{N} \langle w_{ik}\rangle_\infty \overline{x_k}^\mu - \theta_i)\right] \overline{x_j}^\mu \quad j \in V_i \tag{31}$$

The latter result can be rewritten as

$$p\sigma_j^2 \langle w_{ij}\rangle_\infty = -\sum_{k\in V_i} (A_i)_{jk} \langle w_{ik}\rangle_\infty + B_{ij}, \quad j \in V_i \tag{32}$$

where we abbreviated

$$\sigma_j^2 = \frac{1}{p}\sum_{\mu=1}^{p}\overline{(x_j - \overline{x_j}^\mu)^2}^\mu$$
$$= \frac{1}{p}\sum_{\mu=1}^{p}\left[\overline{(x_j^2)}^\mu - (\overline{x_j}^\mu)^2\right] \tag{33}$$

and where we split up the sum over all k in a sum over V_i and a sum over its complement V_i^C:

$$(A_i)_{jk} := \sum_{\mu=1}^{p}\overline{x_j}^\mu\overline{x_k}^\mu, \qquad i = 1,\ldots,N; \quad j,k \in V_i \tag{34}$$

$$B_{ij} := \sum_{\mu=1}^{p}[\kappa(2\overline{x_i}^\mu - 1) - (\sum_{k\in V_i^C}\langle w_{ik}\rangle_0\overline{x_k}^\mu - \theta_i)]\,\overline{x_j}^\mu, \qquad i = 1,\ldots,N; \quad j \in V_i \tag{35}$$

Note that the matrix A_i is a symmetric matrix, the dimension of which equals the number of indices in V_i, *i.e.*, the number of adaptable afferent synaptic connections of neuron i. In the matrix B, we could write $\langle w_{ik}\rangle_0$ rather than $\langle w_{ik}\rangle_\infty$, since $\langle w_{ik}\rangle_0 = \langle w_{ik}\rangle_\infty$ for $k \in V_i^C$. It is easy to solve the equation (32). First, rewrite it as

$$\sum_{k\in V_i}[(D_i)_{jk} + (A_i)_{jk}]\,\langle w_{ik}\rangle_\infty = B_{ij} \tag{36}$$

where D_i is the diagonal matrix

$$(D_i)_{jk} := p\sigma_j^2\delta_{jk}, \qquad j,k \in V_i \tag{37}$$

The matrix $D_i + A_i$ is non-singular, and can be inverted. Inserting the explicit form of B_{ij} (35), we then find

$$\langle w_{ij}\rangle_\infty = \sum_{k\in V_i}(D_i + A_i)^{-1}_{jk}\sum_{\mu=1}^{p}[\kappa(2\overline{x_i}^\mu - 1) - (\sum_{l\in V_i^C}\langle w_{il}\rangle_0\overline{x_l}^\mu - \theta_i)]\,\overline{x_k}^\mu \quad j \in V_i \tag{38}$$

where we used that D_i and A_i are symmetric matrices. In the usual treatments of noiseless recurrent neural networks ($\sigma_j = 0$ for all j), one finds for the $w_{ij}(\infty)$ the so-called pseudo-inverse solution [3], [4], which reads, in our notation,

$$w_{ij}^{PI} = w_{ij}(0) + \sum_{\nu,\mu=1}^{p}(C_i^{-1})^{\mu\nu}[\kappa(2\xi_i^\mu - 1) - (\sum_{k=1}^{N}w_{ik}(0)\xi_k^\mu - \theta_i)]\xi_j^\nu \quad j \in V_i \tag{39}$$

where C_i^{-1} is the inverse of the correlation matrix $C_i^{\mu\nu} = \sum_{k\in V_i}\xi_k^\mu\xi_k^\nu$. Apparently, our result (38) is not a simple generalization of the standard result for noiseless recurrent neural networks. Note that the usual pseudo-inverse solution (39) depends on the initial values $w_{ij}(0)$ of all the weights, whereas our solution (38) depends only on $w_{ij}(0)$ for $j \in V_i^C$ and not on the initial value $w_{ij}(0)$ of the changing weights ($j \in V_i$). Apparently, a little bit of noise completely wipes out the effect of the initial state of changing connections, since the result (38) is true for any value of the noise unequal zero.

In the limit that all σ_j (33) vanish, the set of equations (36) becomes underdetermined for $p < N$, since the matrix A_i is then singular. Hence, the solution (38) does not exist for a noiseless net. Explicitly, this can be seen as follows. Let us suppose

that the p average patterns $\{\overline{x_1}^\mu, \overline{x_2}^\mu, \ldots, \overline{x_p}^\mu\}$, $(\mu = 1, \ldots, p)$, span a p-dimensional vector space. Then, for $r > p$ there are coefficients α_{rl} such that

$$\overline{x_r}^\mu = \sum_{l=1}^{p} \alpha_{rl} \overline{x_l}^\mu \tag{40}$$

for all $\mu = 1, \ldots, p$. It follows that every column $(A_i)_{jr}$, (i fixed, j a running index of V_i and r a fixed number larger than p) is a linear combination of the first p columns of $(A_i)_{js}$ (i fixed, j a running index of V_i and s smaller than or equal to p). Consequently, the matrix A_i has a vanishing determinant, and is not invertible. Therefore, in case the average squared deviation (33) would vanish, the unique solution (38) would not exist. The fact that for vanishing variances σ_j our set of equations for the final weights is under-determined has been mentioned already in the introduction, in the text under equation (4).

In [9] the occurrence of the average squared deviation (33) has been overlooked. This enabled the authors to solve the Master Equation (20) in the usual way. By means of the so-called Gauss-Seidel procedure they obtained a modified version of the usual pseudo-inverse solution for the connections, rather than the expression (38).

4. Intermediate values for the weights

Since our approach was simply based on the assumption of convergence of the $\langle w_{ij}\rangle_n$ for $n \to \infty$, we had no knowledge of the intermediate values of the weights $\langle w_{ij}\rangle_n$ for finite n. However, we can predict the evolution of the weights through an iterative procedure. If we repeat the derivation in section 3, starting from (20) in stead of (21), we find

$$\langle w_{ij}\rangle_{n+1} = \langle w_{ij}\rangle_n + \frac{\eta_i}{p} \sum_{\mu=1}^{p} \kappa(2\overline{x_i}^\mu - 1)\overline{x_j}^\mu - \frac{\eta_i}{p} \sum_{\mu=1}^{p} \Big(\sum_{k=1}^{N} \langle w_{ik}\rangle_n \overline{x_k}^\mu - \theta_i \Big) \overline{x_j}^\mu - \eta_i \sigma_j^2 \langle w_{ij}\rangle_n \quad j \in V_i \tag{41}$$

In the limit $n \to \infty$, equation (41) implies (31), provided that the weights $\langle w_{ij}\rangle_n$ converge.

Using the relation (41), one can find, by numerical iteration, the quantities $\langle w_{ij}\rangle_n$ for any n, given the starting values $\langle w_{ij}\rangle_0$. Hence, we can verify numerically that the $\langle w_{ij}\rangle_n$ are independent of these starting values. Moreover, one can study the convergence of the learning procedure. In order to do so, one must make a particular choice for the probability distribution $p^\mu(\boldsymbol{x})$, which, up to now, was left unspecified. For our choice [see (58)], this distribution will depend on a so-called noise parameter b $(0 \le b \le 1)$, such that $\overline{x_j}^\mu = \xi_j^\mu$ and $\sigma_j^2 = 0$ if the noise parameter b vanishes $(b = 0)$. Through the parameter b we can tune the amount of noise during the learning process. Numerical calculations show that the $\langle w_{ij}\rangle_n$ do indeed converge in the limit $n \to \infty$, for arbitrary b, including $b = 0$, if η_i is small enough. Interestingly, convergence times to the final values (38) diverge for a decreasing noise parameter b (*i.e.*, $b \to 0$), but the time of convergence drops to a small value if $b = 0$ (see figure 1), indicating that something peculiar happens in this limit.

In other words, if one demands existence of the solution (38), one may choose b arbitrarily small, but not zero, and convergence to the solution is faster for larger

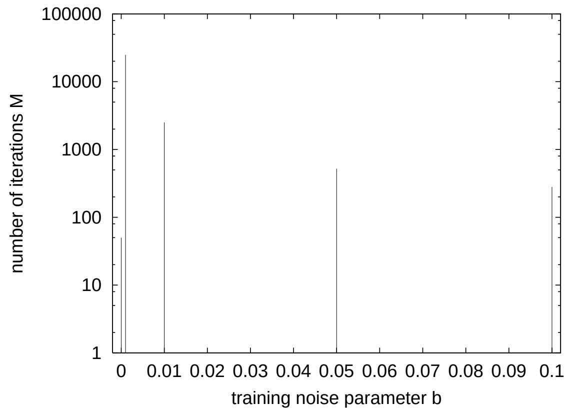

Figure 1. *Convergence time as a function of the training noise parameter.* The number of iterations M of the recursive formula (41) is plotted as a function of the training noise parameter b. The number M is determined with the help of the criterion that it is the smallest value of n for which the condition $\sum_j |\langle w_{ij}\rangle_n - \langle w_{ij}\rangle_\infty| < 0.01$ is satisfied for a fixed value of i. The network consists of $N = 128$ neurons, the number of patterns is $p = 16$. Furthermore, we chose $\theta_i = 0$ for all i and $\kappa = 1$. The learning rate is $\eta_i = 0.25$ for all i. The smaller b, the more iterations are needed to obtain the exact final value $\langle w_{ij}\rangle_\infty$. However, for $b = 0$, M drops to 50. In this case, of course, $\langle w_{ij}\rangle_\infty = w_{ij}^{PI}$.

values of b. If one puts b to zero in the iterative application of (41), one observes rapid convergence of the weights, to the pseudo-inverse values (39). Maybe surprisingly, these values have no continuous relation with the values for finite b, despite the fact that the expressions (41), the difference equations that determine the weights w_{ij}, do depend continuously on the σ_j^2, and, hence [see equation (61) below], on b. In view of the difference in the solutions $\langle w_{ij}\rangle_\infty$ for the cases $b = 0$ [eq. (38)] versus $b \neq 0$ [eq. (39)], this is obvious: there cannot be a continuous relationship between them, since the pseudo-inverse solution (39) depends on all the initial values $w_{ij}(0)$, whereas our solution (38) is independent of the initial values of the weights $w_{ij}(0)$ for $j \in V_i$.

In the next section we investigate whether our final solution for the weights corresponds to the storage of patterns in a stable way.

5. Stability

It is well-known that a neural net with fixed weights w_{ij} (in our case this will be after the learning phase) and deterministic neuron dynamics evolves, in the course of time, to limit cycles of finite length n ($n = 1, 2, \ldots, 2^N$). Cycles with $n = 1$, or fixed points, are of particular interest in neural network theory. If a pattern $\boldsymbol{\xi}^\mu$ is a fixed point of the dynamics of a neural net for a given set of weights w_{ij}, the stability coefficients (3) or (23) are positive for all i, in which case the system remains in the pattern $\boldsymbol{\xi}^\mu$, *i.e.*, the system is stable [10]. Besides the fact that $\gamma_i(\boldsymbol{\xi}^\mu)$ is a measure for the size of the basin of attraction of fixed point $\boldsymbol{\xi}^\mu$ [11], it is plausible that it is also a measure that determines to what extent the network state $\boldsymbol{x}$ remains in a neighborhood of $\boldsymbol{\xi}^\mu$ when the deterministic evolution of this neuron state $\boldsymbol{x}$ is replaced by a stochastic version of this evolution. In order to get an idea of the effectiveness of the learning process

discussed in the preceding section, we will therefore consider the expectation value of the stability coefficients (3) in the limit $n \to \infty$:

$$\langle \gamma_i(\boldsymbol{\xi}^\mu) \rangle_\infty = (2\xi_i^\mu - 1)(\sum_{j=1}^{N} \langle w_{ij} \rangle_\infty \xi_j^\mu - \theta_i) \tag{42}$$

Once a set of patterns $\{\boldsymbol{\xi}^\mu\}$ is known, these quantities can be explicitly calculated with the help of the expression (38). In this section we will attempt to derive, analytically, an approximation of (42) by averaging over sets of patterns $\{\boldsymbol{\xi}^\mu\}$ with given mean activity a. If, however, we would calculate the average of $\langle w_{ij} \rangle_\infty$ over these patterns directly, we would lose all dependence on neuron indices j and pattern indices μ, such that the correlations with the ξ_j^μ would disappear. We therefore have to take a different route.

An approximation for the expectation value (42) is

$$\langle \gamma_i(\boldsymbol{\xi}^\mu) \rangle_\infty \approx (2\xi_i^\mu - 1)(\langle \overline{h_i}^\mu \rangle_\infty - \theta_i) \tag{43}$$

where $\overline{h_i}^\mu$ is the membrane potential of neuron i averaged with respect to $p^\mu(\boldsymbol{x})$:

$$\overline{h_i}^\mu = \sum_{j=1}^{N} w_{ij} \overline{x_j}^\mu \tag{44}$$

In fact, the approximation (43) would be exact if $\overline{x_j}^\mu$ would be equal to ξ_j^μ for all j, *i.e.*, in the limit that the probability function is such that $\overline{x_j}^\mu$ equals ξ_j^μ, for all j. The average potential occurring in (43) can be found from (31). Indeed, multiplying by $\overline{x_j}^\mu$ and summing with respect to $j \in V_i$, we find from this equation:

$$\sum_{j \in V_i} \overline{x_j}^\mu \langle w_{ij} \rangle_\infty p\sigma_j^2 = \sum_{\nu=1}^{p} \Big[\kappa(2\overline{x_i}^\nu - 1) - (\sum_{k=1}^{N} \langle w_{ik} \rangle_\infty \overline{x_k}^\nu - \theta_i) \Big] \sum_{j \in V_i} \overline{x_j}^\nu \overline{x_j}^\mu \tag{45}$$

where we also used (33).

The average square deviation σ_j^2 occurring in this equation depends on the neuron j. In this article we will consider the case in which all neurons have the same standard deviation σ_j

$$\sigma_j = \sigma \qquad j = 1, \ldots, N \tag{46}$$

i.e., the probability $p_j^\mu(x_j)$ is supposed to be such that the uncertainty to find neuron j in a state ξ_j^μ is the same for all neurons of the neural net. Using the identity

$$\sum_{j \in V_i} \overline{x_j}^\mu \langle w_{ij} \rangle_\infty p\sigma^2 = \Big(\sum_{j=1}^{N} \overline{x_j}^\mu \langle w_{ij} \rangle_\infty - \sum_{j \in V_i^C} \overline{x_j}^\mu \langle w_{ij} \rangle_\infty \Big) p\sigma^2 \tag{47}$$

we find from (45)

$$\Big(\langle \overline{h_i}^\mu \rangle_\infty - \sum_{j \in V_i^C} \langle w_{ij} \rangle_0 \overline{x_j}^\mu \Big) p\sigma^2 = \sum_{\nu=1}^{p} [\kappa(2\overline{x_i}^\nu - 1) + \theta_i - \langle \overline{h_i}^\nu \rangle_\infty] \sum_{j \in V_i} \overline{x_j}^\nu \overline{x_j}^\mu \tag{48}$$

where we used $\langle w_{ij} \rangle_0 = \langle w_{ij} \rangle_\infty$ for $j \in V_i^C$. An alternative form for (48) reads

$$\sum_{\nu=1}^{p} [p\sigma^2 \mathbb{1} + C_i]^{\mu\nu} \langle \overline{h_i}^\nu \rangle_\infty = \sum_{\nu=1}^{p} f_i^\nu C_i^{\nu\mu} + g_i^\mu \tag{49}$$

where $\mathbb{1}$ is the $p \times p$ unit matrix and where $C_i^{\mu\nu}$, the correlation matrix for averaged neuron states, is defined by

$$C_i^{\nu\mu} := \sum_{j \in V_i} \overline{x_j}^{\nu} \overline{x_j}^{\mu} \tag{50}$$

Furthermore, we abbreviated

$$f_i^{\mu} = \kappa(2\overline{x_i}^{\mu} - 1) + \theta_i \tag{51}$$

$$g_i^{\mu} = p\sigma^2 \sum_{j \in V_i^C} \langle w_{ij} \rangle_0 \overline{x_j}^{\mu} \tag{52}$$

Multiplying both sides of the matrix equation (49) by the inverse of the (symmetric) matrix occurring on its left-hand side we obtain the solution

$$\langle \overline{h_i}^{\mu} \rangle_{\infty} = \sum_{\nu,\lambda=1}^{p} f_i^{\nu} C_i^{\nu\lambda} [(p\sigma^2 \mathbb{1} + C_i)^{-1}]^{\lambda\mu} + \sum_{\nu=1}^{p} g_i^{\nu} [(p\sigma^2 \mathbb{1} + C_i)^{-1}]^{\nu\mu} \tag{53}$$

Once a particular probability distribution $p^{\mu}(\boldsymbol{x})$ of patterns centered around $\boldsymbol{\xi}^{\mu}$ is given, we can evaluate f_i^{ν} and $C^{\mu\nu}$, and, hence, via (53), the average stability coefficient (43).

In contrast to our expression (38) for the expectation value of the final value of the weights $\langle w_{ij} \rangle_{\infty}$, the result for the expected average potential $\langle \overline{h_i}^{\mu} \rangle_{\infty} = \sum_j \langle w_{ij} \rangle_{\infty} \overline{x_j}^{\mu}$ does exist for vanishing σ. This is clear, already, from (53), in which the existence of the inverse $(p\sigma^2 \mathbb{1} + C_i)^{-1}$ does not depend on the presence of the extra term $p\sigma^2 \mathbb{1}$ as long as the average patterns $\overline{\boldsymbol{x}}^{\mu}$ are linearly independent, since then $C_i^{\mu\nu}$ is invertible. Using (38) and (51), and assuming $\langle w_{ij} \rangle_0 = 0$ for all $j \in V_i^C$ we may write the average potential $\langle \overline{h_i}^{\mu} \rangle_{\infty}$ as

$$\langle \overline{h_i}^{\mu} \rangle_{\infty} = \sum_{\nu} \sum_{k,j \in V_i} (D_i + A_i)^{-1}_{jk} \overline{x_j}^{\mu} \overline{x_k}^{\nu} f_i^{\nu} \tag{54}$$

Comparing this to (53) with (57), we obtain the identity

$$\sum_{k,j \in V_i} (D_i + A_i)^{-1}_{jk} \overline{x_j}^{\mu} \overline{x_k}^{\nu} = \sum_{\lambda} C_i^{\mu\lambda} [(p\sigma^2 \mathbb{1} + C_i)^{-1}]^{\lambda\nu} \tag{55}$$

Hence, though the matrix $(D_i + A_i)^{-1}$ occurring in (54) itself does not exist for $b = 0$, the above combination clearly does: it reduces to $\delta^{\mu\nu}$, as we see from the right hand side for $\sigma = 0$, implying that in the limit of vanishing noise

$$\langle \overline{h_i}^{\mu} \rangle_{\infty} = f_i^{\mu} \tag{56}$$

which is already clear from (53) and is equivalent to —the average of— eq. (2). Thus, although the values of the weights themselves do not have a continuous relation with the values corresponding to the pseudo-inverse solution, the average values for the membrane potentials, and, therefore, of the stability coefficients, do.

In the following we suppose that $w_{ij} = 0$ for all $j \in V_i^C$. This corresponds to a neural net in which all existing connections are of adaptable strength and the only connections with constant strength are the non-existing connections. For $j \in V_i^C$, we then have $w_{ij}(0) = 0$, and, hence, $\langle w_{ij} \rangle_0 = 0$, implying that

$$g_i^{\mu} = 0 \tag{57}$$

Let us choose the probability distribution

$$p_j^{\mu}(x_j) = (1 - b)\delta_{x_j,\xi_j^{\mu}} + b\delta_{x_j,1-\xi_j^{\mu}} \qquad (j = 1, \ldots, N) \tag{58}$$

which fulfills (28), and from which $p^\mu(\boldsymbol{x})$ follows by the prescription (27). The noise parameter b is a probability ($0 \leq b \leq 1$). More specifically, for given μ ($\mu = 1, 2, \ldots, p$), $1-b$ is the probability that the activity x_j of neuron j equals that of the pattern ξ_j^μ. We suppose that b is small compared to unity. As follows from (58), the noise parameter b is related to the width of the distribution of input patterns around each pattern $\boldsymbol{\xi^\mu}$. We can immediately calculate the average neuron state (30) associated with the distribution (58)

$$\overline{x_j}^\mu(\xi_j^\mu) = (1-b)\delta_{1,\xi_j^\mu} + b\delta_{1,1-\xi_j^\mu} \tag{59}$$

the coefficient (51)

$$f_i^\mu = (2\xi_i^\mu - 1)(1-2b)\kappa + \theta_i \tag{60}$$

as well as the average squared deviation (33)

$$\sigma^2 = b(1-b) \tag{61}$$

The fact that σ^2 is j-independent is a consequence of the particular choice (58) for $p_j^\mu(x_j)$, *i.e.*, of the fact that all neurons j are supposed to have the same uncertainty to be in state ξ_j^μ. Hence, the supposition (46) is satisfied.

Let us suppose that the probability that $\xi_j^\mu = 1$ is a, for each j independent of any other neuron index k, and, hence, that the probability that $\xi_j^\mu = 0$ is $(1-a)$, for all of the patterns $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p$. We can now use this to arrive at an estimate value for the average potential of neuron i, eq. (53), which is exact in the limit of $p \to \infty$, for $\alpha = p/N$ fixed, and smaller than 1. From (50) we find, for $\mu \neq \nu$,

$$\begin{aligned} C_i^{\mu\nu} \approx \sum_{j \in V_i} \big\{ & a^2 \overline{x_j}^\mu(1)\overline{x_j}^\nu(1) + a(1-a)\overline{x_j}^\mu(1)\overline{x_j}^\nu(0) + \\ & (1-a)a\overline{x_j}^\mu(0)\overline{x_j}^\nu(1) + (1-a)^2\overline{x_j}^\mu(0)\overline{x_j}^\nu(0)\big\} \end{aligned} \tag{62}$$

while for $\mu = \nu$ we get

$$C_i^{\mu\nu} \approx \sum_{j \in V_i} \{a\overline{x_j}^\mu(1)^2 + (1-a)\overline{x_j}^\mu(0)^2\} \tag{63}$$

Defining the dilution d as the average fraction of neurons from which an arbitrary neuron does not have an incoming connection, each neuron has on the average $N(1-d)$ incoming connections. Hence, using (59), we find from (62) and (63)

$$\begin{aligned} C_i^{\mu\nu} \approx N(1-d)\{ & a(1-a)(1-2b)^2\delta^{\mu\nu} \\ & + [a^2(1-b)^2 + 2ab(1-a)(1-b) + (1-a)^2b^2]\} \end{aligned} \tag{64}$$

We thus have achieved that the correlation matrix $C_i^{\mu\nu}$ for an N neuron net has been expressed in parameters typical for the network, namely the dilution d, the mean activity a and the noise b. An alternative way to write (64) is

$$C_i^{\mu\nu} \approx l\delta^{\mu\nu} + m \tag{65}$$

where l and m are shorthand for combinations of the typical network parameters a, b and d

$$\begin{aligned} l &:= N(1-d)a(1-a)(1-2b)^2 \\ m &:= N(1-d)[a^2(1-b)^2 + 2ab(1-a)(1-b) + (1-a)^2b^2] \end{aligned} \tag{66}$$

With (65), the matrix occurring in (53) can be cast into the form

$$(p\sigma^2\mathbb{1} + C_i)^{\mu\nu} \approx (p\sigma^2 + l)\delta^{\mu\nu} + m \tag{67}$$

The inverse of a p-dimensional matrix A with elements $A^{\mu\nu} = x\delta^{\mu\nu} + y$ is given by the matrix A^{-1} with elements

$$(A^{-1})^{\mu\nu} = [\delta^{\mu\nu}(x + py) - y]/x(x + py) \tag{68}$$

From (65), and (68) applied to (67), we find

$$\sum_{\lambda} C_i^{\nu\lambda}[(p\sigma^2 \mathbb{1} + C_i)^{-1}]^{\lambda\mu} \approx \frac{l[l + p(\sigma^2 + m)]\delta^{\mu\nu} + mp\sigma^2}{(l + p\sigma^2)[l + p(\sigma^2 + m)]} \tag{69}$$

Substituting this result, together with (57), into the expression (53), yields for the average potential of neuron i in pattern μ the expression

$$\langle \overline{h_i}^{\mu} \rangle_\infty \approx \frac{\{l[l + p(\sigma^2 + m)] + mp\sigma^2\} f_i^\mu + mp\sigma^2 \Sigma_{\nu\neq\mu} f_i^\nu}{(l + p\sigma^2)[l + p(\sigma^2 + m)]} \tag{70}$$

The sum over the f_i^ν occurring in this expression can be calculated with (60),

$$\sum_{\nu\neq\mu} f_i^\nu \approx (p - 1)[(2a - 1)(1 - 2b)\kappa + \theta_i] \tag{71}$$

where we used that the average value of the $p - 1$ neuron activities ξ_i^μ can be approximated by a, the average activity of the net. We can now write down the final result for the stability parameters (43), which is a function of the network parameters d, a and b, the number of patterns p, the number of neurons of the net N, and the neuron properties κ and θ_i:

$$\begin{aligned}\langle \gamma_i(\xi_i^\mu) \rangle_\infty \approx\ & \frac{\{l[l + p(\sigma^2 + m)] + mp\sigma^2\}[(1 - 2b)\kappa + \theta_i(2\xi_i^\mu - 1)]}{(l + p\sigma^2)[l + p(\sigma^2 + m)]} \\ & + \frac{mp\sigma^2(p - 1)[(2a - 1)(1 - 2b)\kappa + \theta_i](2\xi_i^\mu - 1)}{(l + p\sigma^2)[l + p(\sigma^2 + m)]} - \theta_i(2\xi_i^\mu - 1)\end{aligned} \tag{72}$$

Note that for $\sigma^2 = 0$ we immediately recover $\langle \overline{h_i}^{\mu} \rangle_\infty = f_i^\mu$ and $\langle \gamma_i(\xi_i^\mu) \rangle_\infty = \kappa$, as we should, from the equations (70) and (72) respectively.

The final average stability coefficient of neuron i takes two different values respectively, depending on whether $\xi_i^\mu = 1$ or $\xi_i^\mu = 0$.

In figure (2) we plotted this quantity for a chosen average activity $a = 0.5$, as a function of b. It is clear that the stability coefficient can be expected to remain positive. In the same figure we plotted the actual values of $\langle \gamma_i(\xi_i^\mu) \rangle_\infty$, as obtained by choosing randomly a set of patterns $\{\boldsymbol{\xi}^\mu\}$ with given mean activity $a = 0.5$, and using (42) with (38) for the calculation.

The difference between the curves is evident, and indicates that we must be careful not to overestimate the accuracy of our result as an indication for $\langle \gamma_i(\xi_i^\mu) \rangle_\infty$. In fact, in a large region, the storage of undisturbed patterns is better than our estimate suggests by a factor 2 to 3, as can be concluded from the figure. With this in mind, we may assume that after the noisy learning process, the patterns $\boldsymbol{\xi}^\mu$ are indeed fixed points under the deterministic network dynamics for a small noise parameter b.

6. Retrieval and basins of attraction

In this section we address the question what happens to the average size of the basins of attraction if noiseless learning (training parameter $b = 0$) is compared to noisy learning ($b \neq 0$). After the network has been trained with patterns $\boldsymbol{x}$ with noise b ($b \neq 0$), we numerically check the retrieval capacity of the net by presenting patterns with noise

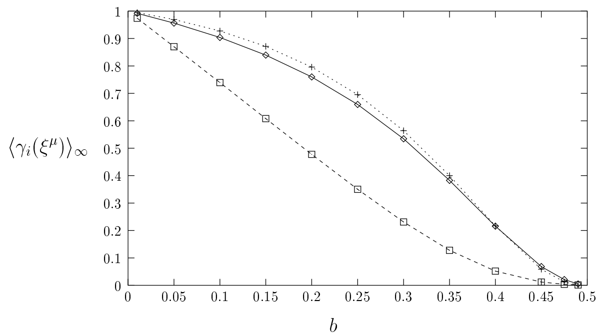

Figure 2. *Average stability as a function of the training noise parameter b.* The average, over all neurons i ($i = 1, 2, \ldots, 256$), and all patterns $\boldsymbol{\xi}^\mu$ ($\mu = 1, \ldots, 32$), of a —diluted— neural net (dilution $d = 0.2$), of the stability coefficients $\gamma_i(\xi_i^\mu)$, with threshold potentials $\theta_i = 0$ for all i, as a function of the noise parameter b. The curves with symbols (+) and ($\Diamond$), for $\xi_i^\mu = +1$ or $\xi_i^\mu = 0$, respectively, correspond to a numerical simulation, and should be compared to the curve with symbol ($\Box$), obtained from the approximative expression (72). The approximation itself is rather poor. However, it gives an indication, as the various curves show. Note that the appearance of one single curve in the approximated case is due to the choice $a = 0.5$ and $\theta_i = 0$.

$b^\star$, *i.e.*, patterns chosen according to a probability distribution of the form (58), in which b has been replaced by $b^\star$. The presented patterns evolve under deterministic parallel dynamics $x_i(n+1) = \Theta(h_i(n) - \theta_i)$. The attempt to retrieve a pattern is successful if the network state $\boldsymbol{x}$ runs into a fixed point equal to the clean, undistorted pattern $\boldsymbol{\xi}^\mu$ of which a noisy version was the initial state. The result can be read off from figure 3. Since the curves obtained via noisy learning lie above the curve with noiseless learning, the basins of attraction are, apparently, enlarged in the presence of noise during the learning stage.

The result is in agreement with earlier studies by Gardner et al [5], and Wong & Sherrington [6], [7], [8].

In [5], like in our case, noise is added to patterns during a training stage. However, the algorithm is of a different kind, because it includes an error-mask, *i.e.*, the weights are updated if and only if, upon presentation of a noisy pattern during the learning stage, the membrane potential h_i has the wrong sign. In this way, if the learning algorithm converges, retrieval of patterns for which the amount of noise is equal to that of the training patterns, is guaranteed.

In [6], [7] and [8], various retrieval properties of a neural network are discussed. It is argued that optimizing (by finding the optimal weights) the overlap of a noisy pattern with its corresponding training representative after one retrieval step is, in fact, a way of noisy training [6]. The optimal network is sought for via a replica-calculation that minimizes a cost-function, thus optimizing the first step retrieval. No explicit learning rule is used in these articles. An explicit expression for the final values of the weights is not given.

Our approach is different from those discussed above, in the sense that we start from an explicit learning rule, which is biologically acceptable: it is derived from the principle that energy cost for synaptic adaptation is minimal [4]; it is a function of local variables; it does not contain error masks; neurons are assumed to be noisy.

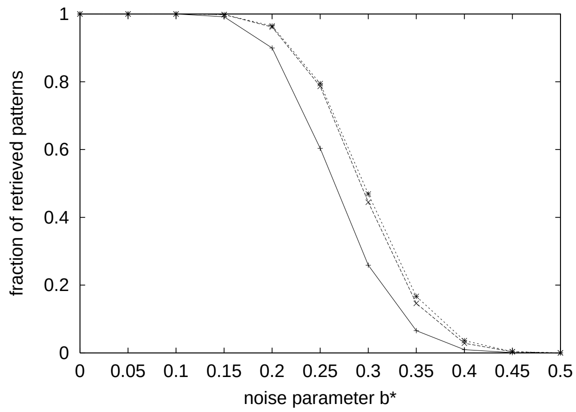

Figure 3. *Fraction of retrieved patterns for networks trained with various values of the training noise parameter b, as a function of the noise $b^\star$ in the patterns to be retrieved.* The average fraction of retrieved patterns for a network of $N = 128$ neurons with thresholds $\theta_i = 0$. The values of the dilution and the activity are $d = 0.2$ and $a = 0.5$. The network has been trained with $p = 32$ patterns. Starting from an initial pattern with noise $b^\star$, the network evolves under parallel deterministic dynamics. The pattern $\boldsymbol{\xi}^\mu$ is said to be retrieved if the network state overlap with this pattern, $m^\mu := N^{-1}\sum_{i=1}^N(2\xi_i^\mu - 1)(2x_i - 1)$, is equal to 1. At most 10 dynamical time steps are applied for every initial pattern. The various curves ($+$), ($\times$) and ($*$) correspond to $b = 0$, $b = 0.05$ and $b = 0.1$ respectively. The value of b for which the size of the basins of attraction is maximal —for the values of b chosen in this figure— is $b = 0.1$. The curve with $b = 0$ (noiseless training) lies below the curves corresponding to learning with noise.

Though in our network the basins of attraction are not optimal (it was not our goal to optimize the basins of attraction), we do have an explicit learning algorithm as well as a final expression for the expectation value of weights.

7. Conclusion

We have shown that learning with noise leads to final values for the weights w_{ij} which are different from those found in the corresponding situation without noise. Surprisingly, the solution for the values of the weights w_{ij} of a noisy system in the limit of vanishing noise, does *not* converge to the values of the solution of the system without noise.

Moreover, in a system without noise the values of the final weights depend on the initial values of all weights, whereas in a noisy system the initial values of the changing weights, the w_{ij} for $j \in V_i$, are wiped out in the course of time.

Our noisy trained networks have larger basins of attraction than noiselessly trained networks. This is in agreement with earlier findings in the literature. The exact dependence of the retrieval properties on various parameters, such as the mean activity a and the memory load $\alpha = p/N$ is still to be elucidated.

Acknowledgment

The authors are indebted to Wouter Kager for carefully reading this manuscript and suggesting some improvements.

References

[1] Müller B, Reinhardt J and Strickland M T 1995 *Neural Networks: An Introduction* (Berlin: Springer)
[2] Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359.
[3] Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949.
[4] Heerema M and van Leeuwen W A 1999 *J. Phys. A: Math. Gen.* **32** 263–86.
[5] Gardner E J, Stroud N and Wallace D J 1989 *J. Phys. A:Math. Gen.* **22** 2019–30.
[6] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** L175–82.
[7] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** 4659–72.
[8] Wong K Y M and Sherrington D 1993 *Phys. Rev. E.* **47** 4465–82
[9] Heerema M and van Leeuwen W A 2000 *J. Phys. A: Math. Gen.* **33** 1781–95.
[10] Kinzel W and Opper M 1991 *Models of Neural Networks* ed Domany E *et al* (Berlin: Springer) p 152
[11] Gardner E 1988*J. Phys. A: Math. Gen* **21** 257–70.